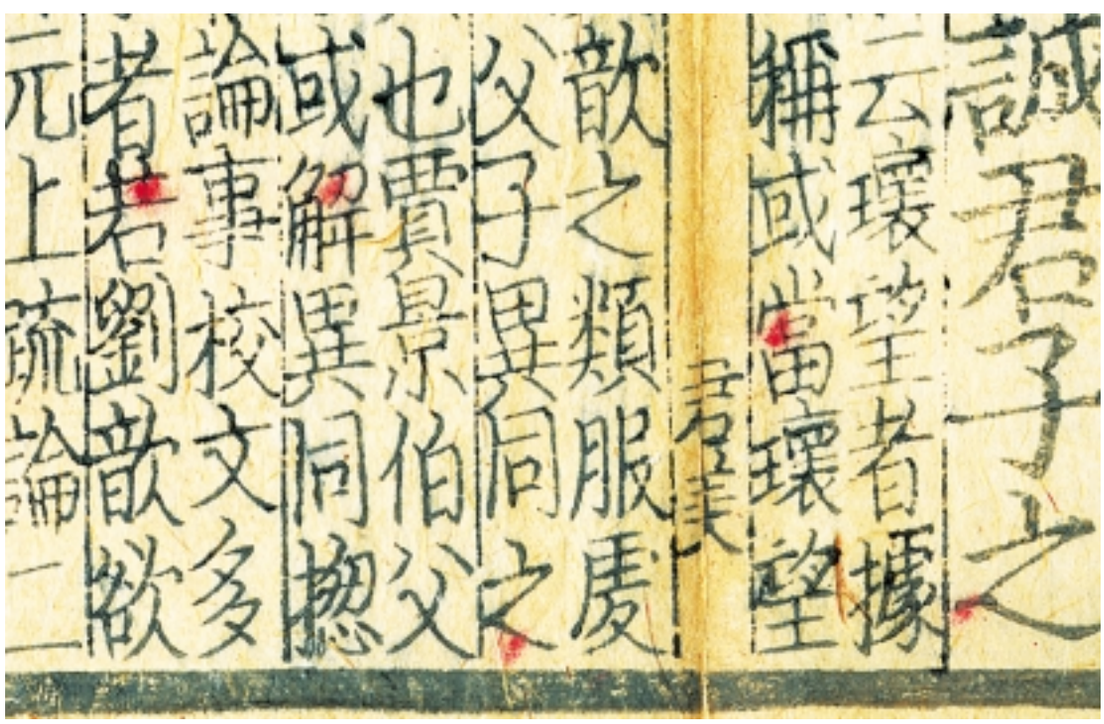
漢字と情報 No.5

漢字と情報

No. 5
2002・10

京都大学人文科学研究所　Documentation and Information Center for Chinese Studies (DICCS)
附属漢字情報研究センター　Institute for Research in Humanities, Kyoto University



● 全国漢籍データベースの現況
● 会読・電子テキスト・XML
● 人文研アーカイブス(5)
　『監本附音春秋穀梁傳註疏』

漢字と情報 2

全国漢籍データベースの現況

高田時雄

漢字情報研究センターの新しい事業として全国漢籍データベースの構築に着手してから約一年半が経過した。準備段階の時日も加えればさらに倍の時間がかかっている。長くもないが，昨今の時間概念からすれば決して短いとも言えない。振り返ればその間には各種の思いがけない困難に直面したこともあった。それらのあるものは現在でもなお未解決のままだが，幸い関係各位の熱心な協同によって，総体としては今のところまずまず順調に発展しつつある。小文ではこのデータベースの現況を簡単に報告しておきたいと思う。

さて，このデータベースはすでにウェブ上で公開されていて（http : // www. kanji. zinbun. kyoto-u. ac. jp / kanseki /），既に利用しておられる方も多いに違いない。現段階では京都大学人文科学研究所，東京大学東洋文化研究所，滋賀大学教育学部，鹿児島大学，立命館大学，京都産業大学の所蔵漢籍データを収録し*，そのレコード数はすでに10万以上に上っている。今年度（平成14年度）は以上に加えて，東北大学，広島大学，千葉県立中央図書館，三康図書館，神戸市外国語大学，高知大学，実践女子大学，新潟大学などのデータを入力作業中で，今年度末か或いは来年度の早い時期には全国データベースに総合される手筈になっている。

データベースの実質的な作成主体は当センターであるが，その作成を支持し各種の問題点を議論する場として全国漢籍データベース協議会が設置されている。この協議会は，国立情報学研究所，東京大学東洋文化研究所附属東洋学研究情報センター，そして我々京都大学人文科学研究所附属漢字情報研究センターの三者が幹事機関となって組織された。後二者は，ともにそれぞれ東洋学文献センターがその前身であり，国内における漢籍資料の保存・整理・公開に中心的役割を果たしてきたという経緯がある。過去二三年の間に相次いで現在の名称のセンターに改組されたが，今後は漢籍の全国データベースの構築に，情報学研究所とともに積極的に取り組んでいきたい。協議会は年一回の頻度で総会が開催されることになっていて，2001年，2002年の３月にすでに２回行われた。その概要は，http : // kanji. zinbun. kyoto-u. ac. jp / kansekikyogikai / を御覧いただきたい。

データベースの作成は年度ごとの科学研究費成果公開費に拠っているため，必ずしも安定した経費の来源が保証されている訳ではないが，2001，2002年度は幸い関係当局のご理解によって申請が認められ，仕事を継続することが可能となっている。来年度以降も引き続き御支持をお願いするのはもちろんだが，年度計画を策定する上で漢籍所蔵機関（図書館）の積極的な参加が重要な鍵となる。参加していただける機関が前もって決定していないと，経費の申請そのものも出来なくなるわけで，その意味でも漢籍を所蔵しておられる機関のご協力を是非ともお願いしたいところである。

ところで，わが国ではこれまで大量の「漢籍目録」が作られてきている。こうした各所蔵機関ごとに作成された目録の数はおそらく数百に上るであろうと思われる。それぞれが貴重な努力の結晶であることは言うまでもないが，それを一つ一つ検索することは余りにも手間が掛かりすぎる。また冊子本の目録を基礎として，各機関で単独のデータベースを作成されているところも，ここ何年かの間に比較的よく見かけるようになってきた。データベースの形になれば，迅速かつ多様な検索が保証されるために，非常に便利であることはいうまでもないが，しかし個々の機関のデータベースを別個に検索しなければならないという不便さ

*必ずしもここに挙げた各機関，各大学のすべての漢籍を収録しているわけではなく，特殊文庫として一括されているものだけである場合もある。また技術的な問題があって，現時点ではごく一部しか収録されていないものも含んでいるが，これらは今年度内にはすべて解決される見込みである。

は一向に解決されない。したがって現在要請されているのは，出来るだけ早く全国的な規模の連合目録データベースを構築することである。そしてそれを実現するためには，これまでに出版された各機関ごとの漢籍目録データを単一のフォーマットに則って入力し統合することが先決である，と我々は考えた。事実，現在行っている作業は冊子の形になった漢籍目録データを入力し校正することが中心になっている。（そのフォーマットと入力例とは，上記データベース協議会のページに見える。）

冊子目録は出来ていないけれども，カード目録なら完備しているという場合にも，冊子に準じる扱いで，決して入力が不可能ではないので，ご相談いただければ幸いである。また漢籍は所蔵しているけれども，整理が行き届いていないか，まったく整理が出来ていないという場合もあり得るだろう。その時には，当センターが文部科学省と共催で行っている漢籍担当職員講習会に是非御参加いただきたい（http：// www. kanji. zinbun. kyoto-u. ac. jp / courses / index. html. ja）。中国書誌学の初歩から漢籍データベース作成の実務までが懇切に解説される。また漢籍データベースをコンピュータで作成するためのツールの開発も行っており，ディスプレー上で必要な項目を逐次入力していけば，データベースに取り込むことが可能な形で保存できるソフトウェアが既に完成し，本年度の漢籍講習会からその実習を始める。このソフトは講習会の参加者だけでなく，協議会のページからもダウンロードが可能である。

漢籍データベースと NII の Webcat との関連はどうなっているのか，という質問をよく受ける。この点については，すでに情報学研究所と当センターの間で研究会が組織され，技術的な検討が開始されていることを報告しておくべきであろう。その結果，細部の問題点は若干残るものの，全国漢籍データベースのデータを Webcat に流すことはほぼ可能だという結論に達しつつある。したがって全国漢籍データベースに提供して頂いた漢籍の所蔵データは，もし所蔵機関の同意があれば Webcat に総合することが近い将来可能となる。

それでは Webcat で事は足りるのではないかという意見もあるかも知れない。しかし漢籍データベースは，漢籍の分類法としてオーソドックスな四部分類を保持し，かつ漢籍の特徴である叢書の子目を階層的に表現できるという，Webcat にない特徴をもっている。更には今後，画像サンプルを付け加える予定になっていて，これは各エディションの判別に絶大な効力を発揮する筈である。

普段，日本国内で利用されている方々はお気付きではないかもしれないが，このデータベースでは漢字コードに Unicode（UTF-8）を採用しているために，中国でも台湾でも，或いは欧米でも，世界中のどこであっても，その地域で普通に行われている漢字入力法を用い，その地域で用いられている漢字コードによって検索できるように作られている。わが国の JIS はもちろんのこと，GB でも BIG5 でもすべて検索が可能なのである。そういう意味で初めから国際的な利用を念頭に置いているので，機会があれば是非とも諸外国の関係者に宣伝して頂きたい。

眼を外に向ければ，ここ何年かのあいだに，中国や台湾さらには欧米をも巻き込んで，国際的な古籍データベースへの取り組みが活発化している。我々の全国漢籍データベースも，日本国内に止まらず国際的な連携を強化していかねばならないが，これは今後の課題である。（人文科学研究所教授）

会読・電子テキスト・XML

岩井茂樹

東方学研究部の共同研究班では，全文電子テキストをつくったうえで，あるいはつくりながら会読をすすめることが普通になった。解釈に迷う語句があれば，まずはその文献中に用例をもとめ，文脈と語句との関係をさぐる。電子テキストの検索は，ここでその威力を発揮する。また，冊子体の索引刊行を考えても，語句の切りだしを自動化できないという泣き所はあるものの，逐字索引であれば簡単なプログラムで電子テキストから派生させることができる。さらには，校定本文や訳注の作成も，電子テキストを土台として作業がおこなわれる。

つくる本人にとっては，本文の情報にくわえて，巻，章，節，段落をきる程度にとどめ，さしたる構造も附加情報ももたない単純なテキストで充分かもしれない。ディスクに保存された複数のファイルから目的の語句をさがす用途であれば，これで事たる。grep の正規表現をもちいた検索をするばあいは，夾雑物のない単純なテキストが好ましい。

しかし，研究班の参加者に検索手段を提供し，また電子テキストを作業の土台として利用してもらうとなると，話は違ってくるだろう。当初のファイルが誤りを多くふくむことは避けられない。会読がすすむにつれて，誤脱が訂正され，句読が検討され，改善されてゆく。複数の版本にもとづく校定情報がくわえられることもある。各人にファイルを配布する方法では，更新のたびごとに，再配布などの手間が避けられない。また，あちこちで訂正や附加の作業が並行しておこなわれることもありうる。それを再び一つにまとめるのは容易ではない。作業過程にあるファイルは一元的な管理のもとにおきたいわけである。

その一方，電子テキストを土台とした作業は，

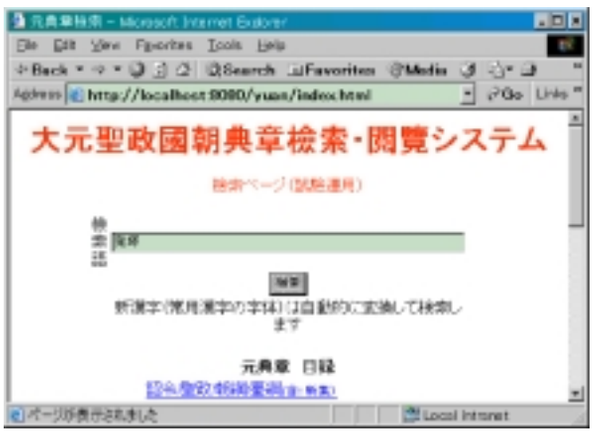


Fig. 1　検索ページ　「発塚」を全文より検索

参加者が個別に分散しておこなう。研究室や自宅から利用できる検索・閲覧手段を提供し，かつ最新であることが保証されたテキストから必要部分を自分の作業環境のなかにとりこめるようなネットワーク上のシステムが望ましい。

さらに，校定本文や検索手段などを研究成果として公開するのであれば，電子媒体であれ，紙媒体であれ，テキストの構造にもとづく表現をあたえなければならない。すると，テキストの構造や種々の付加情報を機械と人間の両方が解釈できるように印づけしたテキストを，会読の作業とともにつくりあげることが考慮されてよい。

XML＝eXtensive Mark-up Language は，電子テキストの印づけと処理の方法として有力な候補である。本研究所の Christian Wittern 氏は，禅の語録を対象とするプロジェクトのなかに XML をとりいれているし，かの『四庫全書』全文検索版も，表面には見えないが，XML を使っているという。分野はことなるけれども判例の電子化に XML を使おうという動きもあるらしい。XML 関連の参考書や記事は豊富であり，習得もむつかしくはない。

しかし，いくつかの懸念があった。XML 文書には，多くのタグが挿入されている。語句がタグによって分断され，複数の要素にまたがることもある。これを正しく検索するしくみが用意されているだろうか？　また，XML 文書はツリー構造

における親子関係や順番についての情報を保持したノードに分節されてメモリ上に展開される。したがって，メモリ上のサイズは，元の XML 文書の数倍にもなるらしい。ノードをたどりながらその中の文字列を走査するのであるから，単純なファイル走査と比べると，処理時間が長くなる。どれほどの速度で検索できるのだろうか？

あちこち調べてみたがよくわからない。そこで，入力作業を終えたばかりの『大元聖政國朝典章』六十巻　附『新集至治條例』不分巻を対象として，Web サーバ上でその検索と閲覧を提供するシステムを試作することにした。この『元典章』全文テキストは本研究所の金文京氏の主唱のもとにつくられ，入力は櫻井智美女史（研究支援推進員／非常勤研究員）と劉暁氏（中国社会科学院歴史研

Fig. 2　検索結果　文中に「発塚」を含む記事の題目を表示

究所）の尽力による。

『元典章』は全体で100万字ほど，タグつけは，筆者一人でおこなうことを考え，極力すくなくした。DTD という様式にしたがって，XML 文書の文書定義（構造の指示）を記述しておくと，不適切なタグつけは，検証プログラムが指摘してくれる。記事ごとの ID や注の番号付与は，後述する XSLT のスクリプトを書いて自動化した。巻ごとに１文書として，全体では69のファイル，計3. 4メガバイトの XML 文書ができあがった。文字セットについては，XML は内部処理を Unicode でおこなうので，UTF-8というエンコーディングを採用した。

こうして XML 文書が用意できると，つぎは，これを検索・閲覧するしくみをつくる段である。さきにのべたように，訂正や付加情報をくわえていくべき作業途上の文書であるから，サーバ上で管理する。その上で，ネットワーク経由で利用者の要求をうけつけ，要求に応じたデータを送りかえしてブラウザ上に表示する。閲覧の場合，XML 文書から動的に生成された目次をたどりながら見たい記事をクリックすると，その記事だけが抽出され HTML に変換されて返ってくる。サーバ側で利用者からの要求をうけつけて，応答する処理の流れは，サーブレットという Java プログラムが処理する。要求に応じて動的に各種 HTML 文書を生成する手順は，XSLT ファイルに記述しておく。XSLT は XPath というノードを指示するための式や関数群とともにもちいて，XML 文書の構造を変えたり，表示方法を指定したりするスクリプト言語である。

全文検索の対象となるのは，計69の XML 文書であるが，これを一つづつ検索するよりも，メモリ上にすべてを展開しておき，一気に検索するほうが効率的である。一方，XML 文書自体は，巻ごとに分けたまま管理するほうが都合がよい。数メガバイトもの長大文書にまとめてしまうと，訂正や加工のさいにあつかいにくい。検索時には大きなかたまりがよく，それ以外の時は巻ごとの小さな文書がよい。相反する要求を満たす解はいずこに？

あれこれ考えているうちに，「外部解析対象実体参照」を使えばよいということに気がついた。まず，文書定義のなかで，巻ごとの文書（これが解析対象実体に相当する）について，それぞれ任意の名前で参照することを定義する。

〈!ENTITY ydz01 SYSTEM “01¥YuanDZ01. xml”〉
〈!ENTITY ydz02 SYSTEM “02¥YuanDZ02. xml”〉
　　（以下略）

上のように，69の文書に結びついた参照を定義したうえで，元典章全体の XML 文書は，実質的な中身は空っぽ，ルート要素である Vols 要素のなかに，「& 名前 ;」という書式で参照をならべておく。

〈Vols〉
　　&ydz01 ;
　　&ydz02 ;
　　（以下略）
〈/Vols〉

つまり，殻だけの XML 文書を用意しておき，メモリに展開するときに，計69の文書から中身を注入するわけだ。各巻の文書に訂正や追加があっても，殻にあたる XML 文書を再読みこみするだけで，部分と全体の同期がとれる。作業途上にある文書の管理方法として理想的である。

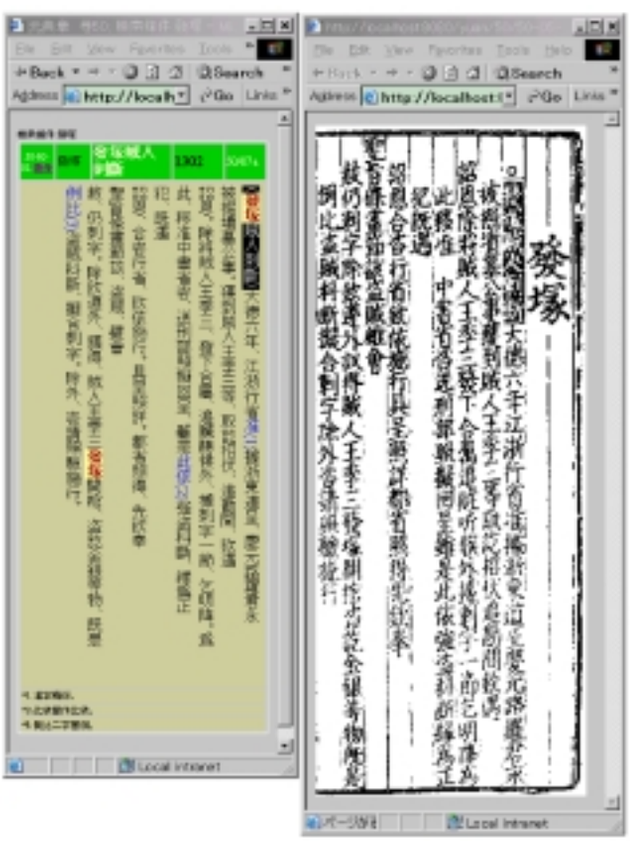


Fig. 3　記事の表示　縦書きのテキストを表示　左上角の「畫像」をクリックすると刊本の画像を表示する

さて，肝腎の XSLT＋XPath による全文検索の速度はどうだろうか。サーバ起動後はじめてサーブレット経由で検索用の XSLT をうごかすときには，メモリ上に全体のノードツリーを構築する処理がある。しかし，いったん構築されたノードツリーは，サーバが動いているかぎりメモリ上に保持される。二回目以降は，検索文字列をサーバに送ってから，該当する記事の標題一覧が返ってくるまで４秒弱。検索対象の長大さを考えるならば，上々の速度であろう。筆者が使った XSLT 処理エンジンは，SAXON と命名された Java 言語で書かれたものである。Tomcat というサーバもやはり Java である。Java のシステムは処理速度が遅いのではないかと恐れていたが，杞憂であった。

ちなみに，４秒弱というのは，PentiumIII 866MH のマシンでサーバを稼働させたばあいの数値である。試しに，同じサーバを Athlon 1800＋（約1. 5GH）のマシンでうごかしたところ，全文検索の速度は２秒以下に短縮された。筆者はあまりに速すぎる CPU の速度は，半ばこけおどしだと考えていたのだが，おおいに認識をあらためた次第である。

（人文科学研究所教授）

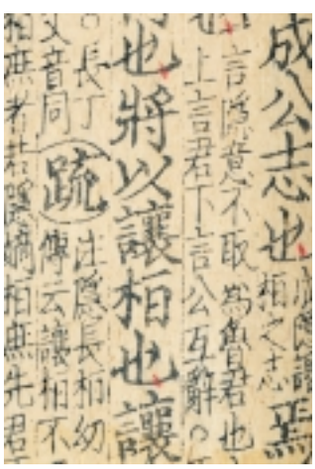
人文研アーカイブス（５）

監本附音春秋穀梁傳註疏

二十卷

晉　范甯　集解　唐　陸德明
音義　唐　楊士勛　疏

元刊本　明修



金鑲玉装，27.3×16.8，原料紙縦22.9，匡廓内18.8×12.8。白口，双黒魚尾，左右双辺，有界，半葉10行17字，小字双行23字。版心「谷亢幾（丁付）」，上象鼻に大小字数，下象鼻に刻工名を刻す。各葉左上欄外に耳格，「隱元年」等とあり。「桓」字を欠筆する箇所多し。君美・以徳・以・天易・住・住郎・伯寿・以清など三十余の刻工標記がある。

首に，「監本附音春秋穀梁傳註疏序」を付す。巻第一首題は「監本附音春秋穀梁註疏隱公卷第一」とし，その下に割書小字で「起元年／盡三年」と刻す。各巻ほぼ同様であるが，巻第五至第十四第十七至第二十は「附音」の二字を欠く。

本書は，いわゆる十行本の注疏合刻本の一である。注疏合刻形式の経書の刊行は宋代より行われていたが，現存するものは少ない。古くは，現存の十行本の多くが宋刊本とされていたが，今日ではほとんどが宋刊本を覆刻した元刊本とされ，宋刊本は足利学校遺蹟図書館蔵建安劉叔剛刊本等ごく少数にとどまる。

ここに掲載した本所蔵本は，研究所草創期に田中文求堂より購入したもので，目録では今日まで宋刊本としているが，他本と比較検討するに，元代の覆刻で，明代の印刷にかかるものとするのが妥当であろう。吉川幸次郎「東方文化研究所善本提要」（『東方學報　京都』第十册第二分）に，

　此亦十行本而無一補頁……

とするが，巻第九第十八葉は，線黒口で，匡廓も18.2×12.2とやや小さく，字様も異なり，この葉のみ明代の補刻であろう。伝来をしめす印記や書入は見あたらない。

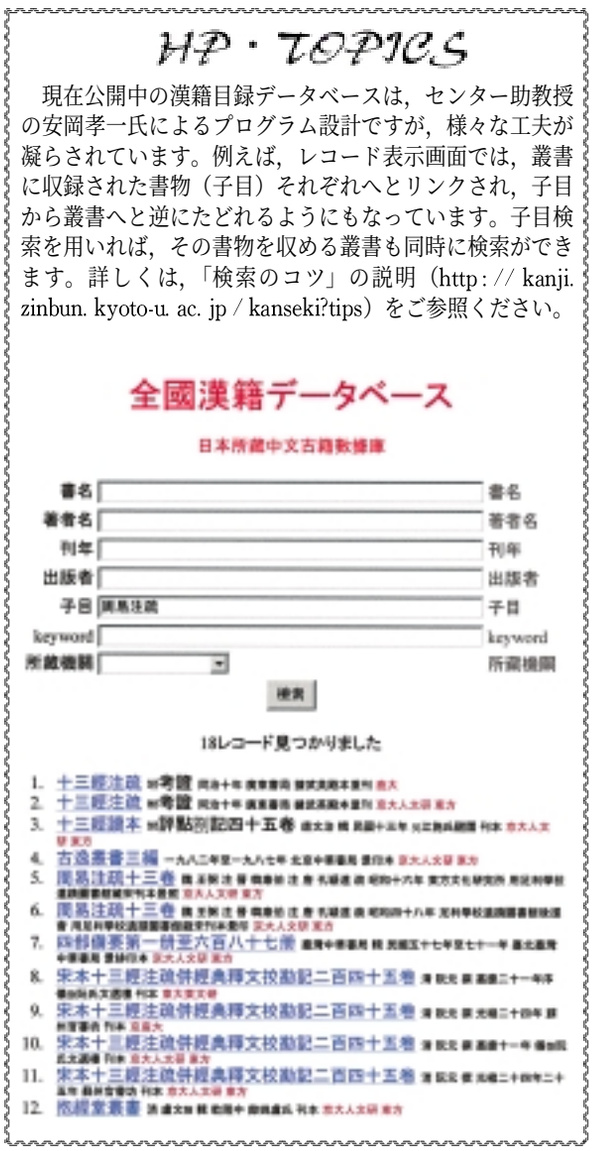
同版が，静嘉堂文庫，中国国家図書館，故宮博物院（台北），国家図書館（台北）などに所蔵されている。

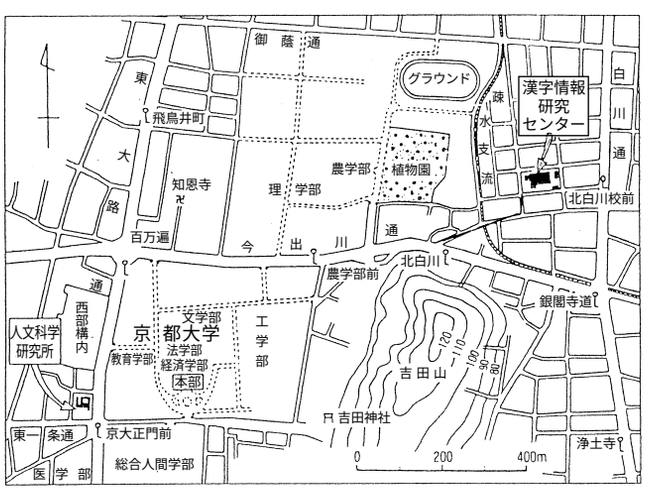
本所には，この穀梁伝以外に，目録で同じく宋刊本とする，注疏合刻十行本の左氏伝と公羊伝があるが，いずれも元刊明修本である。公羊伝は穀梁伝と同時に文求堂から購入しているが，左氏伝は，別途に購入したもので，「莫友芝／圖書印」と「莫印／繩孫」（陰刻）の二顆の朱印記がある。

（センター助手）

HP・TOPICS

現在公開中の漢籍目録データベースは，センター助教授の安岡孝一氏によるプログラム設計ですが，様々な工夫が凝らされています。例えば，レコード表示画面では，叢書に収録された書物（子目）それぞれへとリンクされ，子目から叢書へと逆にたどれるようにもなっています。子目検索を用いれば，その書物を収める叢書も同時に検索ができます。詳しくは，「検索のコツ」の説明（http : // kanji. zinbun. kyoto-u. ac. jp / kanseki?tips）をご参照ください。





【DICCS NEWS】

• 全国漢籍データベースと NACSIS-CAT との相互乗り入れに関して，国立情報学研究所，京都大学附属図書館と本センターが合同で検討を行うことになり，今夏から技術的問題点を討議する漢籍共同研究会を定期的に開催している。

• 文部科学省と本センターとの共催による漢籍担当職員講習会は，これまで初級・中級及び漢籍電算処理の３コースの講習会を行ってきた。ところが，図書整理業務にパソコン利用が一般化し，伝統的な漢籍整理と電算処理を統合的に行うことが求められるようになったので，本年度から新たに初級・中級の２コースを設け，プログラムの大幅な改訂を行うことにした。その詳細は次号に掲載する予定でいる。

なお，本年度受講者は初級が19名，中級が20名（予定），日程は以下の通りである。

初級：10月７日（月）～10月11日（金）

中級：11月11日（月）～11月15日（金）

• 平成14年度教育改善推進費（学長裁量経費）プロジェクト「現代中国に関する教育・研究基盤整備」で，所蔵中国語雑誌の書誌データベース化に着手することになった。類目データベースとリンクさせ，Web 上での公開を計画している。

発行日　2002年10月15日

発行所　京都大学人文科学研究所附属
漢字情報研究センター

〒606-8265 京都市左京区北白川東小倉町47

電話 075-753-6997　FAX 075-753-6999

http://www.kanji.zinbun.kyoto-u.ac.jp/